TALN 2006, Leuven, 10–13 avril 2006

Décodage conceptuel et apprentissage automatique : application au corpus de dialogue Homme-Machine MEDIA

Christophe Servan Frédéric Béchet
LIA, Université d'Avignon
339 chemin des Meinajaries, 84911 Avignon cedex 09
{christophe.servan,frederic.bechet}@univ-avignon.fr

Résumé Cette étude présente les travaux du LIA effectués sur le corpus de dialogue homme-machine MEDIA et visant à proposer des méthodes d'analyse robuste permettant d'extraire d'un message audio une séquence de concepts élémentaires. Le modèle de décodage conceptuel présenté est basé sur une approche stochastique qui intègre directement le processus de compréhension au processus de Reconnaissance Automatique de la Parole (RAP). Cette approche permet de garder l'espace probabiliste des phrases produit en sortie du module de RAP et de le projeter vers un espace probabiliste de séquences de concepts. Les expériences menées sur le corpus MEDIA montrent que les performances atteintes par notre modèle sont au niveau des meilleurs systèmes ayant participé à l'évaluation sur des transcriptions manuelles de dialogues. En détaillant les performances du système en fonction de la taille du corpus d'apprentissage on peut mesurer le nombre minimal ainsi que le nombre optimal de dialogues nécessaires à l'apprentissage des modèles. Enfin nous montrons comment des connaissances *a priori* peuvent être intégrées dans nos modèles afin d'augmenter significativement leur couverture en diminuant, à performance égale, l'effort de constitution et d'annotation du corpus d'apprentissage.

Abstract Within the framework of the French evaluation program MEDIA on spoken dialogue systems, this paper presents the methods proposed at the LIA for the robust extraction of basic conceptual constituents (or concepts) from an audio message. The conceptual decoding model proposed follows a stochastic paradigm and is directly integrated into the Automatic Speech Recognition (ASR) process. This approach allows us to keep the probabilistic search space on sequences of words produced by the ASR module and to project it to a probabilistic search space of sequences of concepts. The experiments carried on on the MEDIA corpus show that the performance reached by our approach is state of the art on manual transcriptions of dialogues. By partitioning the training corpus according to different sizes, one can measure the impact of the training corpus on the decoding performance and therefore estimate the minimal as well as the optimal number of dialogue examples needed. Finally we detail how *a priori* knowledge can be integrated in our models in order to increase their coverage and therefore lowering, for the same level of performance, the amount of training corpus needed.

Mots-clefs : Dialogue Homme-Machine, Reconnaissance Automatique de la Parole, Apprentissage Automatique à base de corpus

Keywords: Spoken dialogue, Automatic Speech Recognition, Corpus-based methods

1 Introduction

Dans les applications de dialogue homme-machine téléphonique, le processus d'interprétation consiste à extraire du message oral des structures conceptuelles. Cette opération ne se résume pas forcément à une analyse de la transcription textuelle du message par une grammaire syntaxico sémantique. Plusieurs considérations étayent cette proposition : d'une part les règles d'interprétation peuvent être contextuelles ; d'autre part, dans le traitement de la parole spontanée, des parties entières du message peuvent être inutiles à la compréhension de celui-ci et l'opération de reconnaissance de concepts peut être un succès même si l'ensemble du message n'est pas complètement engendré par une grammaire. Enfin, la même séquence de mots peut être utile à la reconnaissance de plus d'un concept.

Plusieurs formalismes ont été proposés pour décrire des structures sémantiques. Ils sont essentiellement basés sur les concepts d'entités et de relations. En général les concepts généraux représentant l'interprétation complète d'un message sont obtenus par des opérations de composition sur des concepts élémentaires. Ces concepts sont relativement indépendants du modèle sémantique global utilisé. Ils représentent à la fois les objets sémantiques manipulés par l'application, correspondant à des catégories d'entités nommées telles que les dates, les prix, ou encore les noms propres (ville, hôtel, ...) ; mais aussi les actes dialogiques. C'est sur ces concepts élémentaires que s'est focalisée la campagne d'évaluation MEDIA (programme Technolangue/Evalda), qui consistait à évaluer les capacités d'interprétations de plusieurs systèmes sur un corpus de traces de dialogue homme-machine portant sur un serveur d'informations touristiques.

Cette étude présente les travaux du LIA effectués sur le corpus MEDIA et visant à proposer des méthodes d'analyse robuste permettant d'extraire d'un message audio une séquence de concepts élémentaires. Ces concepts sont les entités utilisées pour construire une interprétation sémantique complète des messages traités. La campagne MEDIA était structurée en deux phases : une phase d'évaluation de la compréhension *hors contexte* et une autre *en contexte*. Dans la première les énoncés sont traités indépendamment les uns des autres, sans aucune information sur le dialogue en cours. Dans la deuxième, les concepts sont enrichis avec les informations contextuelles obtenues lors des précédents tours de dialogue. Nous nous focaliserons dans cette étude sur l'interprétation *hors contexte* des concepts élémentaires.

Ce papier est organisé comme suit : après avoir rapidement présenté la problématique du décodage conceptuel dans le cadre du projet MEDIA, nous présenterons les deux approches principales utilisées pour résoudre ce problème, celle basée sur une analyse syntaxico sémantique et celle qui envisage ce processus comme un processus de traduction automatique. Le paragraphe 2.3 présente le cadre théorique de cette étude puis les différents composants de l'approche proposée sont détaillés dans les paragraphes 3 et 4. Enfin le paragraphe 5 présentera les résultats obtenus par notre approche sur le corpus MEDIA.

2 Décodage conceptuel pour les systèmes de dialogue

Les applications de dialogue homme-machine considérées dans cette étude peuvent être vues comme une interface entre un utilisateur et une base de données. Le but du dialogue est de remplir tous les champs d'une requête qui va être adressée à la base de données. Dans ce cadre les concepts sémantiques de base sont de 3 types : les concepts relatifs au type de la requête ; les

concepts relatifs aux valeurs qui instancient les paramètres de la requête ; et enfin les concepts relatifs à la conduite du dialogue. La campagne d'évaluation MEDIA (Bonneau-Maynard *et al.*, 2005) (programme Technolangue/Evalda) se place dans ce cadre applicatif à travers la simulation d'un système d'accès à des informations touristiques et des réservations d'hôtel. Un corpus de 1250 dialogues a été enregistré par ELDA selon un protocole de *Magicien d'Oz* : 250 locuteurs ont effectués chacun 5 scénarios de réservation d'hôtel avec un système de dialogue simulé par un opérateur humain. Ce corpus a ensuite été transcrit manuellement, puis annoté sémantiquement selon un dictionnaire sémantique de concepts mis au point par les partenaires du projet MEDIA (Bonneau-Maynard *et al.*, 2005). Ce corpus est décrit brièvement dans le prochain paragraphe.

2.1 Le corpus MEDIA

Le dictionnaire sémantique utilisé pour annoter le corpus MEDIA (Bonneau-Maynard *et al.*, 2005) permet d'associer 3 types d'information à un mot ou un groupe de mots :
– tout d'abord une paire attribut-valeur, correspondant à une représentation sémantique *à plat* d'un énoncé ;
– puis un spécifieur qui permet de définir des relations entre les attributs et qui par conséquent peut être utilisé pour construire une représentation hiérarchique de l'interprétation d'un énoncé ;
– enfin une information sur le *mode* attaché à un concept (positif, affirmatif, interrogatif ou optionnel).

n	W^{c_n}	c_n	*mode*	*spécifieur*	*valeur*
0	euh	null	+		
1	oui	réponse	+		oui
2	l'	LienRef-coRef	+		singulier
3	hôtel	BDObjet	+		hotel
4	dont	null	+		
5	le prix	objet	+	chambre	paiement-montant
6	ne dépasse pas	comparatif-paiement	+		inferieur
7	cent dix	paiement-montant-entier	+	chambre	110
8	euros	paiement-devise	+		euro

TAB. 1 – Exemple de message annoté du corpus MEDIA

La table 1 présente un exemple de message annoté du corpus MEDIA. La première colonne correspond au numéro du segment dans le message, la deuxième colonne à la chaîne de mots W^{c_n} porteuse du concept c_n contenu dans la troisième colonne. Les colonnes 4, 5 et 6 contiennent le mode, le spécifieur et la valeur du concept c_n dans la chaîne W^{c_n}. Le dictionnaire sémantique MEDIA contient 83 attributs, auxquels peuvent s'ajouter 19 spécifieurs de relations entre attributs. Le corpus collecté a été découpé en plusieurs lots. Nous utilisons dans cette étude les 4 premiers lots comme corpus d'apprentissage, soit 720 dialogues contenant environ 12K messages d'utilisateurs, et le lot 5 comme corpus de tests contenant 200 dialogues avec 3K messages d'utilisateurs.

2.2 D'un flux de mots vers un flux de concepts

Le traitement de transcriptions automatiques de messages oraux se caractérise par deux phénomènes particuliers : d'une part les phénomènes dus à la parole spontanée tels que les disfluences (hésitations, reprises, auto-corrections, dislocations, incises) ; d'autre part l'absence de structure dans les sorties des systèmes de transcription automatique parole-texte. Cette absence de structure se caractérise par la génération d'un flux de mots sans ponctuation ni découpage en phrase. La seule segmentation généralement opérée repose sur des silences réalisés par le locuteur avec des longueurs supérieures à un seuil fixé. Ces deux phénomènes rendent très difficile toute analyse complète de ce type de message et les processus de compréhension se doivent d'opérer sur des analyses partielles. En dehors de la traditionnelle opposition *méthodes à base de connaissance a priori*/*méthodes à base d'apprentissage automatique*, les modèles développés pour répondre au problème du décodage conceptuel d'un flux de mots peuvent être vus selon deux perspectives :

– soit comme une conséquence d'un processus d'analyse syntaxico sémantique qui va construire une ou plusieurs analyses structurées du message à analyser, dans ce cas les concepts à détecter sont des noeuds dans la ou les structures obtenues ;
– soit comme le résultat d'une opération de traduction automatique qui consiste à transformer une suite de symboles donnée en entrée (les mots) en une autre suite de symboles (les concepts).

Pour illustrer la première famille de méthodes, on peut citer un système comme TINA (Seneff, 1992) basé sur des analyses complètes et partielles ; d'autres systèmes utilisent une analyse robuste incrémentale basée sur une étape de *chunking* comme dans (Antoine *et al.*, 2003) ; (Wang *et al.*, 2002) présente des grammaires hors-contexte syntaxico sémantiques dérivées à partir de patrons génériques.

La deuxième famille de méthode se rapproche du cadre théorique utilisé dans les applications de Reconnaissance Automatique de la Parole (RAP). Dans ce cadre le décodage de parole est vu comme la transmission d'un signal dans un canal bruité. Le but est de décoder le message initial à partir des observations (des paramètres acoustiques dans le cadre de la RAP, des mots dans le cadre du décodage conceptuel) qui ont transité à travers le canal de communication. Cette opération de *traduction* se réalise de manière probabiliste en cherchant l'interprétation C qui maximise la probabilité $P(C|A)$, A représentant la séquence d'observation acoustique. Cette approche, initiée par les travaux de (Levin & Pieraccini, 1995), se retrouve dans de nombreux systèmes de décodage conceptuel tels que (Bonneau-Maynard & Lefevre, 2005).

2.3 Modèle théorique

Le modèle théorique utilisé dans cette étude est basé sur la deuxième approche présentée au paragraphe précédent. Ce choix théorique est basé sur deux considérations :
– les chaînes de mots données en sortie des systèmes de RAP sont généralement produites à l'aide de modèles ayant une portée très faible (modèles bigrammes ou trigrammes), ne garantissant aucune cohérence au delà d'une fenêtre de quelques mots ; pour cette raison il est important de pouvoir garder des hypothèses multiples, sous la forme d'un graphe de mots ;
– même si des travaux précédents ont montré que des analyses syntaxiques peuvent être appliquées à des graphes de mots (Chappelier *et al.*, 1999; Roark, 2002), les disfluences et l'absence de structure dans le flux de parole rendent ces analyses difficiles, limitées le plus

souvent à de multiples analyses partielles ; de fait l'intérêt principal du processus d'analyse, l'obtention d'une analyse fonctionnelle complète de chaque composant du message, est rarement atteint.

Nous noterons C l'interprétation d'un message. C représente une séquence de concepts de base, tels que ceux défini dans le corpus MEDIA. Le décodage conceptuel consiste à chercher la chaîne de concepts $C = c_1, c_2, \ldots, c_k$ maximisant $P(C|A)$, A étant la séquence d'observations acoustiques. En utilisant le même paradigme que celui utilisé en RAP, trouver la meilleure séquence de concepts $\hat{C}$ exprimé par la séquence de mots W à partir de la séquence d'observation acoustique A s'exprime par la formule suivante :

$$P(\hat{C}|A) \approx \underset{C}{\textit{argmax}} \sum_{W} P(A|W,C)P(W,C) \approx \underset{C,W}{\textit{argmax}}\, P(A|W)P(W,C) \qquad (1)$$

La probabilité $P(A|W)$ correspond à la probabilité estimée par les modèles acoustiques pour la chaîne de mots W. $P(W,C)$ est la probabilité jointe d'une chaîne de mots W et d'une chaîne de concepts C. Cette recherche de la meilleure interprétation $\hat{C}$ va être faite dans un graphe de mots produit par le système de RAP pour chaque message traité. La première étape dans cette recherche consiste à transformer ce graphe de mots en un graphe de concepts. Ce processus est présenté dans le paragraphe suivant.

3 D'un graphe de mots vers un graphe de concepts

Dans cette étude les concepts de base composant une interprétation C sont notés c_i. A chaque concept c_i correspond la chaîne de mots W^{c_i} qui supporte le concept et à partir de laquelle la valeur associée (par exemple la date ou encore une information numérique) peut être extraite. L'interprétation d'un message contenant L concepts est représenté à la fois par la séquence $C = \{c_1, c_2, \ldots\ c_L\}$ et par la séquence de chaînes de mots $W^C = \{W^{c_1}, W^{c_2}, \ldots, W^{c_L}\}$. Cette notation est illustré par l'exemple de message donné dans la table 1.

Etant donné que les concepts de base c_i sont des entités simples de taille finie, il est possible de représenter chacun d'entre eux par une grammaire régulière codée sous la forme d'un automate à états finis (ou Finite State Machine, FSM). Ces grammaires sont obtenus de deux manières :

– automatiquement : pour chaque concept c_i, toutes les chaînes de mots W^{c_i} contenus dans le corpus d'apprentissage MEDIA sont regroupés et généralisés en remplaçant certains termes par des symboles non terminaux (termes numériques, éléments de dates, ou encore certaines classes de noms propres) puis ces chaînes de symboles sont regroupés dans un automate A_{c_i} ;
– manuellement : pour certains concepts non spécifiques au corpus MEDIA (tels que les dates ou encore les prix), des grammaires manuelles sont utilisées pour enrichir les automates obtenus automatiquement sur le corpus d'apprentissage. Le paragraphe 5.2 détaille les résultats avec et sans ces grammaires manuelles.

Chacun des automates A_{c_i} ainsi créés est ensuite transformé en transducteur. Le transducteur FSM_{c_i}, pour le concept c_i, accepte la chaîne de mots W^{c_i} sur ses symboles d'entrée et émet le concept c_i sur ses symboles de sortie. Un transducteur FSM_{bck} permettant d'accepter n'importe quelle chaîne de mots en émettant le concept *null* (c'est à dire hors concept) est également défini. Au final, le transducteur $T_{Concept}$ contient l'union de tous ces transducteurs, chacun d'eux pouvant s'enchaîner avec n'importe quel autre, à l'exception du transducteur FSM_{bck} qui ne peut boucler sur lui-même.

Lors du traitement d'un message, le graphe de mots G produit par le système de RAP, lui aussi représenté sous forme de FSM, est composé avec le transducteur $T_{Concept}$, c'est à dire qu'une opération d'intersection est opérée entre les deux automates. Le résultat de cette intersection est un nouveau transducteur $G_{Concept}$ où l'espace de recherche produit par le système de RAP est structuré par rapport aux différents concepts attendus. Les chemins dans $G_{Concept}$ correspondent soit aux suites de concepts C si on considère les symboles de sortie du transducteur, soit aux chaînes de mots W^C en considérant les symboles d'entrée.

Le score d'un chemin dans $G_{Concept}$ est calculé à partir de l'équation 1 qui devient :

$$P(C|A) \approx \underset{W^C \in G_{Concept}}{argmax} \; P(A|W^C)P(W^C, C) \tag{2}$$

Le premier terme est calculé avec les scores acoustiques présents dans le graphe de mots G, l'estimation de la probabilité $P(W^C, C)$ est présentée dans le prochain paragraphe. Le principal intérêt du transducteur $G_{Concept}$ est la possibilité d'obtenir directement les meilleures sequences de concepts pouvant être détectés dans un message en projetant ce transducteur sur ses symboles de sorties, et en énumérant les n-meilleurs chemins. A l'inverse, pour obtenir les k-meilleurs chaînes de mots supports W^C pour une interprétation C, il suffit d'énumérer les k meilleurs chemins de $G_{Concept}$ qui émettent la séquence de concepts C. Ce processus est détaillé dans (Raymond *et al.*, 2006).

4 Choix des meilleures interprétations

La probabilité $P(W^C, C)$ est la probabilité jointe d'avoir à la fois la chaîne de mots $W^C = w_1 w_2 \dots w_n$ et la suite de concepts $C = c_1 c_2 \dots c_l$. A chaque mot w_i on peut associer l'étiquette t_i. Si w_i participe à l'expression du concept c_j alors $t_i = c_j$, si w_i n'est dans aucun concept alors $t_i =$ ***null***. Ainsi, nous aurons : $P(W^C, C) = P\{(t_1, w_1)(t_2, w_2) \dots (t_n, w_n)\} = P(t_{1,n}, w_{1,n})$.

Ce processus est identique à la problématique des étiqueteurs probabilistes, telle qu'on peut la trouver dans (Charniak *et al.*, 1993). En définissant de manière adéquate des termes tels que $t_{1,0}$, ainsi que leurs probabilités, on obtient :

$$P(t_{1,n}, w_{1,n}) = \prod_{i=1}^{n} P(t_i|t_{1,i-1}, w_{1,i-1})P(w_i|t_{1,i}, w_{1,i-1}) \tag{3}$$

De manière à pouvoir estimer ces probabilités, nous faisons les hypothèses de Markov suivantes :

$$P(t_i|t_{1,i-1}, w_{1,i-1}) = P(t_i|t_{i-2,i-1}, w_{i-2,i-1}) \;\; \textbf{\textit{et}} \;\; P(w_i|t_{1,i}, w_{1,i-1}) = P(w_i|t_{i-2,i}, w_{i-2,i-1}) \tag{4}$$

Ainsi nous faisons l'hypothèse que l'étiquette t_i ne dépend que des deux mots et étiquettes précédents. De même le mot w_i ne dépend que des deux mots et étiquettes précédents ainsi que de la connaissance de son étiquette t_i. Nous obtenons l'équation suivante :

$$P(t_{1,n}, w_{1,n}) = \prod_{i=1}^{n} P(t_i|t_{i-2,i-1}, w_{i-2,i-1})P(w_i|t_{i-2,i}, w_{i-2,i-1}) \tag{5}$$

Les étiquettes t_i sont de la forme c_i*_B* ou c_i*_I* si t_i est attachée à un mot exprimant le concept c_i. Le suffixe *_B* signifie que le mot est en début de concept (B pour ***begin***) sinon le suffixe *_I*

est utilisé (I pour *inside*). Si le mot n'appartient à aucun concept, alors $t_i = \textit{null}$. Par exemple la séquence *euh le deux mars à midi* pourra être étiquetée :
`(euh,null)(le,null)(deux,$DATE_B)(mars,$DATE_I)(à,null)(midi,$TIME_B)`
Afin de limiter le phénomène du manque de données, certains mots sont généralisés grâce à un ensemble de symboles non-terminaux correspondant aux chiffres, aux jours de la semaine, aux mois, ... Ainsi l'exemple précédent devient :
`(euh,null)(le,null)($NB,$DATE_B)($MONTH,$DATE_I)(à,null)(midi,$TIME_B)`
Afin d'apprendre les probabilités de l'équation 5, un corpus d'apprentissage contenant des transcriptions de dialogues est nécessaire. Ce corpus doit être manuellement étiqueté en concepts, il est ensuite formaté comme dans l'exemple précédent. Les probabilités de l'équation 5 sont alors apprises selon le critère du maximum de vraisemblance, avec un nécessaire lissage pour les n-grammes non vus. On obtient de cette manière un modèle de langage avec repli qui est utilisé pour estimer la probabilité $P(W^C, C)$ de l'équation 2. Ce modèle de langage est représenté sous forme d'automates grâce à l'ensemble d'outils *FSM/GRM library* d'AT&T (Mohri *et al.*, 2002), il est composé avec le transducteur $G_{Concept}$ pour donner le score final à chaque chemin de $G_{Concept}$.

5 Expériences

Les expériences présentées dans cette étude ont été menées sur le corpus MEDIA en considérant les 83 attributs présentés au paragraphe 2.1. Le mode et les 19 spécifieurs ne sont pas pris en compte ici, ils sont traités dans notre système par le module d'interprétation d'un énoncé en contexte, et ne relèvent donc pas du processus de décodage conceptuel présenté ici.

Le corpus d'apprentissage de 720 dialogues a été découpé en 7 parties contenant respectivement 25, 50, 100, 200, 400, 600 et 720 dialogues. Ces différentes tailles permettent de mesurer l'impact sur les performances de la quantité de dialogues disponibles pour l'apprentissage des modèles. Cet apprentissage concerne deux types de modèles :

– les automates à états finis A_{c_i} représentant les concepts c_i, tels que présentés au paragraphe 3 ;
– le modèle de langage présenté au paragraphe 4 pour estimer la probabilité $P(t_{1,n}, w_{1,n})$.

Les performances sont mesurées par rapport au taux d'erreur sur les paires attribut/valeur. Un concept détecté est considéré comme correct uniquement si l'attribut du concept ainsi que sa valeur normalisée sont corrects d'après la référence. En alignant la chaîne de concepts détectés automatiquement et celle présente dans la référence établie manuellement, on calcule le nombre de concepts corrects C, le nombre de concepts insérés I, le nombre de concepts omis O ainsi que le nombre de concepts substitués S (soit au niveau de l'attribut, soit au niveau de la valeur). Si R est le nombre de concepts de la chaîne référence, le taux d'erreur appelé le *Concept Error Rate* (CER) se calcule selon la formule : $\textit{CER} = \frac{I+O+S}{R} \times 100$.

La campagne MEDIA s'étant d'abord focalisé sur le traitement de transcriptions manuelles, nous présentons ici les résultats obtenus sur ces transcriptions. Le graphe de mots G présenté au paragraphe 2.3 est ainsi réduit à la seule chaîne de mots correspondant à la transcription manuelle de chaque énoncé. Le corpus de tests contient 200 dialogues représentant environ 3000 énoncés et 8500 occurrences de concepts.

5.1 Performances *vs.* Taille de l'apprentissage

Le tableau 2 présente le taux d'erreur sur les concepts (CER) en fonction de la taille du corpus d'apprentissage. La valeur du CER décroît de 46 jusqu'à 24 en augmentant la taille du corpus d'apprentissage, de 25 à 720 dialogues (soit environ 12K énoncés). Notons qu'un CER de 24 est comparable aux performances obtenues par les meilleurs systèmes testés lors de la campagne d'évaluation MEDIA.

Taille corpus	**25**	**50**	**100**	**200**	**400**	**600**	**720**
CER	46.0	43.7	34.6	31.9	26.3	24.2	24.3
correct	61.9%	63.3%	70.9%	74.7%	79.6%	80.8%	81.3%

TAB. 2 – Taux d'erreur sur les concepts (CER) du corpus de test et pourcentage de concepts corrects en fonction de la taille du corpus d'apprentissage (en nombre de dialogue)

La décroissance du taux d'erreur se ralentit fortement après 400 dialogues et se stabilise à 600 dialogues, indiquant qu'une augmentation de la taille du corpus d'apprentissage au delà de 800 dialogues ne se traduirait pas forcément par une amélioration des performances. La taille optimale de corpus pour l'approche présentée ici se situe ainsi autour de 400 dialogues. Le tableau 2 permet aussi d'estimer la taille minimale de corpus nécessaire afin d'obtenir un système présentant des performances acceptables. On remarque ainsi qu'un minimum de 100 dialogues est nécessaire pour avoir un taux d'erreur inférieur à 35 pour un pourcentage de concepts corrects d'environ 70%. Il est intéressant de comparer cette courbe performance/taille de corpus avec celle présentée dans la description de la campagne MEDIA (Bonneau-Maynard *et al.*, 2005), mettant en relation le nombre de dialogues annotés manuellement et la mesure d'accord entre les annotateurs (*IAG*). Cette courbe montre qu'il a fallu environ trois itérations et une centaine de dialogues annotés pour obtenir un taux d'accord inter-annotateurs satisfaisant. Le manuel d'annotations n'a cessé d'évoluer et s'est stabilisé à la fin de l'annotation du premier lot du corpus MEDIA contenant 200 dialogues. On peut ainsi faire le parallèle entre la quantité de corpus nécessaire à la mise au point d'un manuel définissant les concepts d'une application donnée et celle nécessaire à l'apprentissage de modèles statistiques permettant de modéliser ces mêmes concepts.

5.2 Ajout de connaissances *a priori*

Pour remédier au problème du manque de données, surtout dans les premières étapes du processus de développement d'une application où peu de données d'apprentissage sont disponibles, il est tentant d'ajouter des connaissances *a priori* dans les modèles de décodage conceptuel. L'approche présentée dans cette étude permet d'effectuer directement cette intégration : les automates utilisés pour obtenir le graphe de concepts présenté dans le paragraphe 3 ne sont pas stochastiques, ainsi on peut ajouter directement au transducteur $T_{concept}$ d'autres automates représentant des connaissances *a priori* sur les concepts de l'application. Par exemple certaines entités telles que les dates ou des entités numériques se retrouvent dans de nombreuses applications et peuvent être facilement modélisés sous forme de grammaires régulières. Ces grammaires, une fois représentées sous forme d'automates, sont fusionnées avec celles dérivées du corpus d'apprentissage.

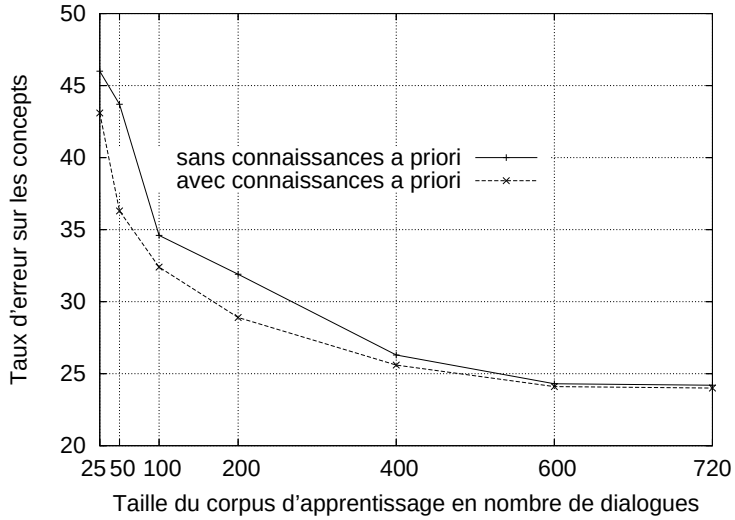

FIG. 1 – Taux d'erreur sur les concepts en fonction de la taille du corpus d'apprentissage, avec et sans connaissances *a priori*

La figure 1 présente les courbes du CER en fonction de la taille du corpus d'apprentissage avec et sans connaissances *a priori*. Ces connaissances ici sont réduites à des grammaires pour les dates et des entités numériques telles que le nombre de chambres ou le nombre de personnes. Comme on peut le voir ces grammaires permettent de palier le manque de données dans les premières phases de l'apprentissage, cet avantage s'estompe à mesure que la couverture du corpus d'apprentissage augmente. Par exemple, grâce à ces connaissances *a priori* les performances du système avec un apprentissage sur 100 dialogues sont identiques à celles obtenus sans ces connaissances avec le double de dialogues d'apprentissage, réduisant ainsi de moitié l'effort d'annotation.

6 Conclusion

Nous avons présenté dans cette étude un modèle de décodage conceptuel, basé sur une approche stochastique, intégré directement dans le processus de Reconnaissance Automatique de la Parole (RAP). L'un des principaux avantages de cette approche est de garder l'espace probabiliste des phrases produit en sortie du module de RAP et de le projeter vers un espace probabiliste de séquences de concepts. Ainsi l'incertitude dans l'interprétation d'un message peut-elle être gardée plus longtemps pour être levée par des niveaux supérieurs d'interprétation intégrant le contexte du dialogue.

Les expériences menées sur le corpus MEDIA montrent que les performances atteintes par notre modèle sont au niveau des performances des meilleurs systèmes ayant participé à la campagne d'évaluation. En détaillant les performances de notre système en fonction de la taille du corpus d'apprentissage on peut mesurer le nombre minimal ainsi que le nombre optimal de dialogues nécessaires à l'apprentissage des modèles. Il est particulièrement intéressant de mettre ces résultats en rapport avec ceux obtenus lors de la constitution du corpus MEDIA sur le nombre de dialogues d'exemples nécessaire afin d'obtenir un manuel d'annotation stable. Les systèmes basés sur une modélisation explicite des connaissances ont besoin d'une telle spécification des concepts. Ainsi ce genre d'étude permet de relativiser l'argument présentant les méthodes à

base d'apprentissage automatique comme nécessairement plus gourmandes en terme de corpus d'apprentissage.

Enfin nous montrons comment des connaissances *a priori* peuvent être intégrées dans nos modèles. Ces connaissances permettent d'augmenter significativement leur couverture et donc de diminuer de manière importante, à performance égale, l'effort de constitution et d'annotation du corpus d'apprentissage.

La prochaine étape dans la campagne MEDIA vise à évaluer l'interprétation d'un énoncé en contexte. Les perspectives de ce travail consistent ainsi à intégrer notre approche dans un tel processus d'interprétation. Des expériences sur des graphes de mots issus du système de RAP du LIA sont également en cours.

Références

ANTOINE J.-Y., GOULIAN J. & VILLANEAU J. (2003). Quand le TAL robuste s'attaque au langage parlé : analyse incrementale pour la compréhension de la parole spontanée. In *TALN'2003*, p. 25–34, Batz-sur-Mer, France.

BONNEAU-MAYNARD H. & LEFEVRE F. (2005). A 2+1-level stochastic understanding model. In *Automatic Speech Recognition and Understanding workshop (ASRU)*, Porto Rico.

BONNEAU-MAYNARD H., ROSSET S., AYACHE C., KUHN A. & MOSTEFA D. (2005). Semantic annotation of the french media dialog corpus. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Lisboa, Portugal.

CHAPPELIER J., RAJMAN M., ARAGUES R. & ROZENKNOP A. (1999). Lattice parsing for speech recognition. In *Proceedings of the 6th conference on Traitement Automatique du Langage Naturel TALN'99, Cargese, Corsica, France*.

CHARNIAK E., HENDRICKSON C., JACOBSON N. & PERKOWITZ M. (1993). Equations for part-of-speech tagging. In *11th National Conference on Artificial Intelligence*, p. 784–789.

LEVIN E. & PIERACCINI R. (1995). Concept-based spontaneous speech understanding system. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, p. 555–558, Madrid, Spain.

MOHRI M., PEREIRA F. & RILEY M. (2002). Weighted finite-state transducers in speech recognition. *Computer, Speech and Language*, **16**(1), 69–88.

RAYMOND C., BÉCHET F., DE MORI R. & DAMNATI G. (2006). On the use of finite state transducers for semantic interpretation. *Speech Communication*, **48,3-4**, 288–304.

ROARK B. (2002). Markov parsing : lattice rescoring with a statistical parser. In *Proceedings of the 40th ACL meeting, Philadelphia*.

SENEFF S. (1992). TINA : A natural language system for spoken language applications. *Computational Linguistics*, **18**(1), 61–86.

WANG Y.-Y., ACERO A., CHELBA C., FREY B. & WONG L. (2002). Combination of statistical and rule-based approaches for spoken language understanding. In *Proc. International Conference on Spoken Language Processing*, p. 609–613, Denver, CO, USA.